

Compositional Data Regression in Insurance with Exponential Family PCA

Guojun Gan^{1a}, Emiliano A. Valdez^{2b}

¹ Department of Mathematics, University of Connecticut, ² University of Connecticut

Keywords: Compositional data, Principal component analysis, Exponential family, Negative binomial regression model

<https://doi.org/10.66573/001c.116404>

Variance

Vol. 17, Issue 1, 2024

Compositional data are multivariate observations that carry only relative information between components. Applying standard multivariate statistical methodology directly to analyze compositional data can lead to paradoxes and misinterpretations. Compositional data also frequently appear in insurance, especially with telematics information. However, such type of data does not receive deserved special treatment in most existing actuarial literature. In this paper, we explore and investigate the use of exponential family principal component analysis (EPCA) to analyze compositional data in insurance. The method is applied to analyze a dataset obtained from the U.S. Mine Safety and Health Administration. The numerical results show that EPCA is able to produce principal components that are significant predictors and improve the prediction accuracy of the regression model. The EPCA method provides a promising useful and flexible tool for actuaries to analyze compositional data.

Address for Correspondence: guojun.gan@uconn.edu

1. INTRODUCTION

Compositional data describe parts of some whole and are commonly presented as vectors of proportions, percentages, concentrations, or frequencies. The sums of these vectors are constrained to be some fixed constant (e.g., 100%) but can be presented in other forms. Due to such constraints, compositional data possess the following properties (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015):

Scale invariance. They carry relative rather than absolute information. In other words, the information does not depend on the particular units used to express the compositional data.

Permutation invariance. Permutation of components of a compositional vector does not change the information contained in the compositional vector.

Subcomposition coherence. Information conveyed by a compositional vector of d components should not con-

tradict that coming from a subcomposition containing $d_*(d_* < d)$ components of the original d components.

These properties are important for understanding the consequences that can make the application of conventional statistical methods restrictive and invalid for compositional data.

Consider the case that such data are multivariate observations that carry relative information between components. Applying standard statistical methodology designed for unconstrained data directly to analyze such compositional features can affect inference leading to paradoxes and misinterpretations. [Table 1](#) shows an example of Simpson's paradox (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015). [Table 1\(a\)](#) shows the number of students in two classrooms who arrive on time and are late, classified by gender. [Table 1\(b\)](#) shows the corresponding proportions of students arriving on time or being late. [Table 1\(c\)](#) shows the aggregated number of students from the two classrooms who arrive on time and are late. The corresponding

a Guojun Gan is an Associate Professor in the Department of Mathematics at the University of Connecticut, where he has been since August 2014. He received a BS degree from Jilin University, Changchun, China, in 2001 and MS and PhD degrees from York University, Toronto, Canada, in 2003 and 2007, respectively. He is also a Fellow of the Society of Actuaries (FSA). His research interests are in the interdisciplinary areas of actuarial science and data science. He has published several books and papers on a variety of topics. According to Google Scholar, his work has been cited more than 4,700 times.

b Emiliano A. Valdez is a Fellow of the Society of Actuaries and has a Ph.D. from the University of Wisconsin in Madison. He joined the University of Connecticut in 2007 but had a brief stint (2013-2015) as professor and director of the actuarial science program at Michigan State University. In recent years, his research work has evolved around the applications of data science and statistical modeling to solve actuarial and insurance problems. The quality of his research has been recognized through awards that include the E. A. Lew Award, the Halmstad Memorial Prize, and the Hachemeister Prize.

(a) Number of students arriving on time and being late, classified by classroom and gender.

Classroom	Gender	On Time	Late
1	M	53	9
1	F	20	2
2	M	12	6
2	F	50	18

(b) Proportion of students arriving on time or being late, classified by classroom and gender.

Classroom	Gender	On Time	Late
1	M	0.855	0.145
1	F	0.909	0.091
2	M	0.667	0.333
2	F	0.735	0.265

(c) Number of students arriving on time or being late, classified by gender.

Gender	On Time	Late
M	65	15
F	70	20

(d) Proportion of students arriving on time or being late, classified by gender.

Gender	On Time	Late
M	0.813	0.188
F	0.778	0.222

Table 1. An example of Simpson’s paradox.

proportions are shown in [Table 1\(d\)](#). From [Table 1\(b\)](#), we see that the proportion of female students arriving on time is greater than that of male students in both classrooms. However, [Table 1\(d\)](#) shows the opposite result: the proportion of male students arriving on time is greater than that of female students. The Simpson’s paradox tells us that we need to be careful about inferring from compositional data. The conclusions made from a population may not be true for subpopulations and vice versa.

Our motivation for studying compositional data regression in insurance comes from the fact that there has been a recent explosion of telematics data analysis in insurance (Verbelen, Antonio, and Claeskens 2018; Guillen et al. 2019; Pesantez-Narvaez, Guillen, and Alcañiz 2019; Denuit, Guillen, and Trufin 2019; Guillen et al. 2020; So, Boucher, and Valdez 2021). In telematics data, we typically observe total mileage traveled by an insured during the year and mileage traveled in different conditions such as at night, over speed, or in urban areas. The distances traveled in different conditions are examples of compositional data. However, compositional data do not receive special treatment in some existing actuarial literature. For example, Guillen et al. (2019) and Denuit, Guillen, and Trufin (2019) selected some percentages of distances traveled in different conditions and treated them as ordinary explanatory variables. Pesantez-Narvaez, Guillen, and Alcañiz (2019) treated compositional predictors as ordinary explanatory variables in their study of using XGBoost and logistic regression to predict motor insurance claims with telematics data. Guillen et al. (2020) also treated compositional predictors as ordinary explanatory variables in their negative binomial regression models. Verbelen, Antonio, and Claeskens (2018) used compositional data in generalized additive models and proposed a conditioning and a projection approach to handle structural zeros in components of compositional predictors; this is one of the few studies in insur-

ance that considers handling compositional data. However, the study does not consider dimension-reduction techniques for compositional data in its regression models. Other areas of actuarial science in which compositional techniques previously have been used include capital allocation problems (see Belles-Sampera, Guillen, and Santolino 2016; Boonen, Guillen, and Santolino 2019) and mortality forecasting (see Kjærgaard et al. 2019).

Compositional data analysis dates back to the end of the nineteenth century when Karl Pearson published an article about spurious correlations (Pearson 1897). The classical method for dealing with compositional data is the logratio transformation method, which preserves the aforementioned properties. As pointed out by Aitchison (2003), the logratio transformations cannot be applied to compositional data with zeros. In insurance data (e.g., telematics data), components of compositional vectors with zeros can be quite common. This sparsity imposes additional challenges to analyze compositional data in insurance. For more information about compositional data analysis, readers are referred to Aitchison and Egozcue (2005); Pawlowsky-Glahn and Buccianti (2011); and Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015).

In this article, we investigate regression modeling of compositional variables (i.e., covariates) by using the exponential family principal component analysis (EPCA) to learn low-dimensional representations of compositional data. Such representations of the compositional predictors extend the classical principal component analysis (PCA) as well as its probabilistic extension via probabilistic PCA, both of which are special cases of EPCA. See Tipping and Bishop (1999). The contributions of this article are summarized as follows:

- It brings the issues of compositional data to actuaries’ attention and provides a framework for incorporating compositional data in actuarial models.

- It presents a detailed case study of compositional data analysis in the insurance context.
- It produces R code that can be used by practitioners and researchers to replicate the analysis and in their future work.

The remaining part of the paper is organized as follows. In Section 2, we present four negative binomial regression models for fitting the data. In Section 3, we provide numerical results of applying the proposed regression models to a mine injury dataset. In Section 4, we conclude the paper with some remarks.

2. METHODS AND MODELS

In this section, we present the models for the compositional data described in the previous section. In particular, we first present the logratio transformation in detail. Then we present a dimension-reduction technique for compositional data. Finally, we present regression models with compositional covariates.

2.1. METHODS FOR COMPOSITIONAL DATA ANALYSIS

The classical method for dealing with compositional data is the logratio transformation method, which preserves the aforementioned properties. Let $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a compositional data set with n compositional vectors. These vectors are assumed to be contained in the simplex

$$\mathbb{S}^d = \left\{ \mathbf{x} \in \mathbb{R}^d : \sum_{j=1}^d x_j = \kappa; \forall j, x_j > 0 \right\},$$

where $\kappa > 0$ is a constant, usually 1. The superscript d in $\mathbb{S}^d$ does not denote the dimensionality of the simplex as the dimensionality of the simplex is $d - 1$.

The logratio transformation method transforms the data from the simplex to the real Euclidean space. There are three common types of logratio transformations (Aitchison 1994): centered logratio transformation (CLR), additive logratio transformation (ALR), and isometric logratio transformation (ILR). CLR scales each compositional vector by its geometric mean; ALR applies logratio between a component and a reference component; and ILR uses an orthogonal coordinate system in the simplex.

Let $\mathbf{x}$ be a column vector (i.e., a $d \times 1$ vector). CLR is defined as

$$\begin{aligned} \text{clr}(\mathbf{x}) &= \log \frac{\mathbf{x}}{g(\mathbf{x})} \\ &= \log(\mathbf{x}) - \frac{1}{d} \sum_{j=1}^d \log(x_j) \\ &= M_c \log(\mathbf{x}), \end{aligned} \quad (1)$$

where $g(\mathbf{x}) = \left(\prod_{j=1}^d x_j \right)^{1/d}$ is the geometric mean of $\mathbf{x}$ and $M_c = I_d - \frac{1}{d} \mathbf{1}_d \mathbf{1}_d^T$ is a $d \times d$ matrix. Here I_d is a $d \times d$ identity matrix and $\mathbf{1}_d$ is a $d \times 1$ vector of ones, i.e.,

$$\begin{aligned} M_c &= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} - \frac{1}{d} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} - \frac{1}{d} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \end{aligned}$$

ALR is defined similarly. Suppose that the d th component of compositional vectors is selected as the reference component. Then ALR is defined as

$$\text{alr}(\mathbf{x}) = \log(\mathbf{x}) - \log(x_d) = M_a \log(\mathbf{x}), \quad (2)$$

where x_d is the d th component of $\mathbf{x}$ and M_a is a $(d - 1) \times d$ matrix given by

$$M_a = (I_{d-1} \quad -\mathbf{1}_{d-1}) = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & -1 \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix}.$$

ILR involves an orthogonal coordinate system in the simplex and is defined as

$$\text{ilr}(\mathbf{x}) = R \cdot \text{clr}(\mathbf{x}), \quad (3)$$

where R is a $(d - 1) \times d$ matrix that satisfies the following conditions:

$$RR^T = I_{d-1}, \quad R^T R = M_c = I_d - \frac{1}{d} \mathbf{1}_d \mathbf{1}_d^T.$$

The matrix R is called the contrast matrix (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015). CLR and ILR have the following relationship:

$$\begin{aligned} R^T \text{ilr}(\mathbf{x}) &= R^T R \text{clr}(\mathbf{x}) = R^T R R^T R \log(\mathbf{x}) \\ &= R^T R \log(\mathbf{x}) = \text{clr}(\mathbf{x}). \end{aligned}$$

Further, we have

$$R \text{clr}(\mathbf{x}) = R R^T \text{ilr}(\mathbf{x}) = \text{ilr}(\mathbf{x}).$$

All three logratio transformation methods are equivalent up to linear transformations. ALR does not preserve distances. The CLR method preserves distances but can lead to singular covariance matrices. The ILR method avoids these drawbacks.

These logratio transformation methods provide one-to-one mappings between the real Euclidean space and the simplex and allow transforming back to the simplex from the Euclidean space without loss of information. The reverse operation of CLR, for example, is given by

$$\mathbf{x} = \text{clr}^{-1}(\mathbf{z}) = \frac{\kappa \exp(\mathbf{z})}{\sum_{j=1}^d \exp(z_j)}, \quad (4)$$

where $\mathbf{z} = \text{clr}(\mathbf{x})$ is the transformation of $\mathbf{x}$ under the CLR method. The reverse operation of ILR is given by

$$\mathbf{x} = \text{clr}^{-1}(R^T \text{ilr}(\mathbf{x})), \quad (5)$$

where clr^{-1} is defined in Equation (4).

2.2. PCA FOR COMPOSITIONAL DATA

In this subsection, we introduce some dimension-reduction methods for compositional data. In particular, we introduce the traditional PCA and the EPCA for compositional data in details. The EPCA methodology has recently been shown to

Table 2. Examples of some convex functions and the corresponding Bregman divergences.

$\varphi(\mathbf{x})$	$D_\varphi(\mathbf{x}, \mathbf{y})$
$\frac{1}{2}\ \mathbf{x}\ ^2$	$\frac{1}{2}\ \mathbf{x} - \mathbf{y}\ ^2$
$\sum_{j=1}^d \exp(x_j)$	$\sum_{j=1}^d e^{y_j}(e^{x_j - y_j} - 1 - x_j + y_j)$

transform compositional data to uncover meaningful structures (Avalos et al. 2018).

The EPCA method is a generalized PCA method for distributions from the exponential family that was proposed and formulated by Collins, Dasgupta, and Schapire (2002) based on the same ideas of generalized linear models. To describe the traditional PCA and the EPCA methods, let us first introduce the φ -PCA method because both the traditional PCA and the EPCA methods can be derived from the φ -PCA method.

We consider a data set $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, consisting of n numerical vectors, each of which consists of d components. For notation purpose, we assume all vectors are column vectors and X is a $d \times n$ matrix.

Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex and differentiable function. Then the φ -PCA aims to find matrices $A \in \mathbb{R}^{l \times n}$ and $V \in \mathbb{R}^{l \times d}$ with $VV^T = I_l$ that minimize the following loss function:

$$\mathcal{L}_{\varphi\text{-PCA}}(X; A, V) = \sum_{i=1}^n D_\varphi(\mathbf{x}_i, V^T \mathbf{a}_i), \quad (6)$$

where $\mathbf{a}_i$ is the i th column of A and D_φ denotes the Bregman divergence:

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T [\nabla \varphi(\mathbf{y})]. \quad (7)$$

Here ∇ denotes the vector differential operator. A Bregman divergence can be thought of as a general distance. Table 2} gives two examples of convex functions and the corresponding Bregman divergences.

For an exponential family distribution, the conditional probability of a value x given a parameter value θ has the following form:

$$\log P(x|\theta) = \log P_0(x) + x\theta - G(\theta), \quad (8)$$

where $P_0(x)$ is a function of x only, θ is the natural parameter of the distribution, and $G(\theta)$ is a function of θ that ensures that $P(x|\theta)$ is a density function. The negative log-likelihood function of the exponential family can be expressed through a Bregman divergence. To see this, let F be a function defined by

$$F(h(\theta)) + G(\theta) = h(\theta)\theta,$$

where $h(\theta) = G'(\theta)$ is the derivative of G . From the above definition, we can show that $f(\theta) = h^{-1}(\theta)$, where $f(\theta) = F'(\theta)$. Hence,

$$\begin{aligned} -\log P(x|\theta) &= -\log P_0(x) - x\theta + G(\theta) \\ &= -\log P_0(x) - x\theta + h(\theta)\theta - F(h(\theta)) \\ &= -\log P_0(x) - F(x) + F(x) - F(h(\theta)) \\ &\quad - (x - h(\theta))\theta \\ &= -\log P_0(x) - F(x) + F(x) - F(h(\theta)) \\ &\quad - (x - h(\theta))f(h(\theta)) \\ &= -\log P_0(x) - F(x) + \varphi_F(x, h(\theta)), \end{aligned}$$

which shows that the negative log-likelihood can be written as a Bregman distance plus two terms that are constant with respect to θ . As a result, minimizing the negative log-likelihood function with respect to θ is equivalent to minimizing the Bregman divergence.

The EPCA method aims to find matrices $A \in \mathbb{R}^{l \times n}$ and $V \in \mathbb{R}^{l \times d}$ with $VV^T = I_l$ that maximize the following loss function:

$$\begin{aligned} \mathcal{L}_{EPCA}(X; A, V) &= \log P(X|A, V) \\ &= \sum_{i=1}^n \sum_{j=1}^d \log P(x_{ji}|\theta_{ji}), \end{aligned} \quad (9)$$

where

$$\theta_{ji} = \sum_{h=1}^l a_{hi} v_{hj}$$

or

$$\Theta = (\theta_{ji})_{d \times n} = V^T A.$$

In other words, the EPCA method seeks to find a lower dimensional subspace of parameter space to approximate the original data. The matrix V contains l basis vectors of $\mathbb{R}^d$, and $\theta_i = (\theta_{1i}, \theta_{2i}, \dots, \theta_{di})^T$ is represented as a linear combination of the basis vectors. The traditional PCA is a special case of EPCA when the normal distribution is appropriate for the data. In this case, $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$.

To apply the traditional PCA to compositional data, we first need to transform them from the simplex into the real Euclidean space by using a logratio transformation method (e.g., the CLR method). When the compositional data are transformed by the CLR method, the traditional PCA has the following loss function:

$$\mathcal{L}_{PCA}(X; A, V) = \frac{1}{2} \|\text{clr}(X) - V^T A\|^2, \quad (10)$$

where $\text{clr}(X) = (\text{clr}(\mathbf{x}_1), \text{clr}(\mathbf{x}_2), \dots, \text{clr}(\mathbf{x}_n))$ is the CLR-transformed data.

Avalos et al. (2018) proposed the following EPCA method for compositional data:

$$\mathcal{L}_{EPCA}(X; A, V) = D_{\exp}(\text{clr}(X), V^T A), \quad (11)$$

where $D_{\exp}(\cdot, \cdot)$ is the Bregman divergence corresponding to the $\exp(\cdot)$ function. With this setup, we employ an optimization procedure to find A and V such that the loss function is minimized.

2.3. REGRESSION WITH COMPOSITIONAL COVARIATES

Compositional data in regression models are a set of predictor variables that describes parts of some whole and is commonly presented as vectors of proportions, percentages, concentrations, or frequencies.

Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015) presented a linear regression model with compositional covariates. The linear regression model is formulated as follows:

$$\hat{y}_i = \beta_0 + \langle \beta, \mathbf{x}_i \rangle_a, \quad (12)$$

where β_0 is a real intercept and $\langle \cdot, \cdot \rangle_a$ is the Aitchison inner product defined by

$$\begin{aligned} \langle \beta, \mathbf{x}_i \rangle_a &= \langle \text{clr}(\beta), \text{clr}(\mathbf{x}_i) \rangle = \langle \text{ilr}(\beta), \text{ilr}(\mathbf{x}_i) \rangle \\ &= \sum_{j=1}^d \log \frac{\beta_j}{g(\beta)} \log \frac{x_j}{g(\mathbf{x})}. \end{aligned}$$

Here $g(\cdot)$ denotes the geometric mean. The sum of squared errors is given by

$$\begin{aligned} SSE &= \sum_{i=1}^n (y_i - \beta_0 - \langle \beta, \mathbf{x}_i \rangle_a)^2 \\ &= \sum_{i=1}^n (y_i - \beta_0 - \langle \text{ilr}(\beta), \text{ilr}(\mathbf{x}_i) \rangle)^2, \end{aligned} \quad (13)$$

which suggests that the actual fitting can be done using the ILR-transformed coordinates, that is, a linear regression can be simply fitted to the response as a linear function of $\text{ilr}(\mathbf{x})$. CLR should not be used because it requires the generalized inversion of singular matrices.

Since the response variable of the mine data set is a count variable, it is not suitable to use linear regression models. Instead, we will use generalized linear models to model the count variable. Common generalized linear models for count data include Poisson regression models and negative binomial models (Frees 2009). Poisson models are simple but restrictive as the Poisson distribution has only one parameter. The negative binomial distribution has two parameters. In this study, we use negative binomial models as they provide more flexibility than do Poisson models.

Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ denote the predictors from n observations. The predictors can be selected in different ways. For example, the predictors can be the ILR-transformed data or the principal components produced by a dimension-reduction method such as the traditional PCA and the EPCA methods. For $i = 1, 2, \dots, n$, let y_i be the response corresponding to $\mathbf{x}_i$. In a negative binomial regression model, we assume that the response variable follows a negative binomial distribution:

$$P(y = j) = \frac{\Gamma(j+r)}{j! \Gamma(r)} p^r (1-p)^j, \quad j = 0, 1, 2, \dots, \quad (14)$$

where $r > 0$ and $p \in (0, 1)$ are parameters and $\Gamma(\cdot)$ denotes the gamma function. The mean and the variance of a variable y are given by

$$E[y] = \frac{r(1-p)}{p}, \quad \text{Var}(y) = \frac{r(1-p)}{p^2}.$$

To incorporate covariates in a negative binomial regression model, we let the parameter p vary by observation. In particular, we incorporate the covariates as follows:

$$\mu_i = \frac{r(1-p_i)}{p_i} = E_i \exp(\mathbf{x}_i' \beta), \quad (15)$$

where r and β are parameters to be estimated and E_i is the exposure. For the mine data set, we use the variable AVG_EMP_TOTAL as the exposure. This is similar to modeling the number of claims in a Property and Casualty data set where the time to maturity of an auto insurance policy is used as the exposure (Frees 2009; Gan and Valdez 2018). The method of maximum likelihood can be used to estimate the parameters.

In this study, we compare four negative binomial models. Table 3 describes the covariates of the four models. In the first model, we use the original compositional components as the covariates. Since the sum of the compositional components is equal to 1, we exclude the first component in the regression model to avoid overparameterization. In the second model, we use the compositional components transformed by the ILR method that is defined in Equation (3). In

Table 3. Description of negative binomial regression models for the mine dataset.

Model	Covariates
NB	TYPE_OF_MINE and the original compositional components
NBILR	TYPE_OF_MINE and the ILR-transformed compositional components
NBPCA	TYPE_OF_MINE and principal components from the traditional PCA method
NBEPCA	TYPE_OF_MINE and principal components from the EPCA method

the third model, we use the first few principal components obtained from the traditional PCA method that is defined in Equation (10). In the fourth model, we use the first few principal components obtained from the EPCA method that is defined in Equation (11).

2.4. VALIDATION MEASURE

To measure the accuracy of regression models for count variables, it is common to use Pearson's chi-square statistic, which is defined as

$$\chi^2 = \sum_{j=0}^m \frac{(O_j - E_j)^2}{E_j}, \quad (16)$$

where O_j is the observed number of mines that have j injuries, E_j is the predicted number of mines that have j injuries, and m is the maximum number of injuries considered over all mines. Between two models, we prefer the model with a lower chi-square statistic.

The observed number of mines that have j injuries is obtained by

$$O_j = \sum_{i=1}^n I_{\{y_i=j\}}, \quad j = 0, 1, \dots, m,$$

where I is an indicator function. The predicted number of mines that have j injuries is calculated as follows:

$$\begin{aligned} E_j &= E \left[\sum_{i=1}^n I_{\{y_i=j\}} \right] = \sum_{i=1}^n E[I_{\{y_i=j\}}] = \sum_{i=1}^n P(\hat{y}_i = j) \\ &= \sum_{i=1}^n \frac{\Gamma(j+\hat{r})}{j! \Gamma(\hat{r})} \hat{p}_i^{\hat{r}} (1-\hat{p}_i)^j, \end{aligned}$$

where $\hat{r}$ and $\hat{p}_i$ are estimated values of the parameters.

From Figure 2, we see that the maximum observed number of injuries is 86. To calculate the validation measure, we use $m = 99$, that is, we use the first 100 terms of O_j s and E_j s. The remaining terms are too close to zero and will be ignored.

In addition to the chi-square statistic, we compare the average observed number of injuries and the predicted number of injuries. To do that, we use the following mean prediction error:

$$ME = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i), \quad (17)$$

where n is the number of observations, y_i denotes the observed number of injuries in the i th mine, and $\hat{y}_i$ denotes the predicted number of injuries in the i th mine. Between

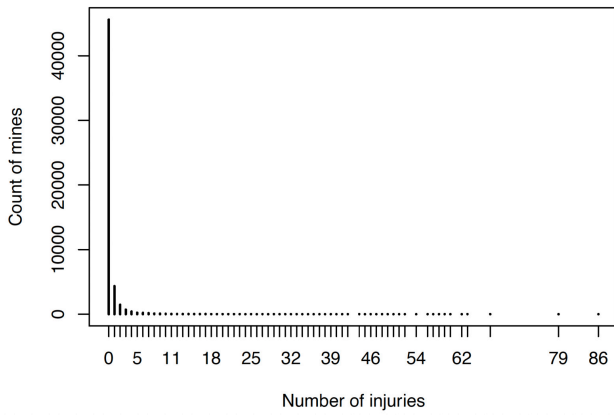


Figure 2. A frequency distribution of the response variable.

two models, we prefer the model with a lower absolute value of the mean prediction error.

3. DATA ANALYSIS

In this section, we demonstrate the performance of the models described in the previous section to a mine injury data set obtained from the Internet. Although our research was motivated by telematics data, we do not have access to real telematics data. As a result, we use this mine injury data set to showcase the application of compositional data analysis in actuarial science and insurance.

3.1. DESCRIPTION OF THE DATA

We use a data set from the U.S. Mine Safety and Health Administration from 2013 to 2016. The data set was used in the Predictive Analytics exam administered by the Society of Actuaries in December 2018.¹ This data set contains 53,746 observations described by 20 variables, including compositional variables. Table 4 shows the 20 variables and their descriptions. Among the 20 variables, 10 of them (i.e., the proportion of employee hours in different categories) are compositional components.

In our study, we ignore the following variables: YEAR, US_STATE, COMMODITY, PRIMARY, SEAM_HEIGHT, MINE_STATUS, and EMP_HRS_TOTAL. We do not use YEAR in our regression models as we do not study the temporal effect of the data. However, we use YEAR to split the data into a training set and a test set. The variables COMMODITY, PRIMARY, and SEAM_HEIGHT are related to the variable TYPE_OF_MINE, which we will use in our models. The variable EMP_HRS_TOTAL is highly correlated to the variable AVG_EMP_TOTAL, with a correlation coefficient of 0.9942. We use only one of them in our models.

Table 5 shows summary statistics of the categorical variable TYPE_OF_MINE. We also show the number of observations in different years. From the table, we see that there are ap-

Table 4. Description of the 20 variables of the mine dataset.

Variable	Description
YEAR	Calendar year of experience
US_STATE	U.S. state where mine is located
COMMODITY	Class of commodity mined
PRIMARY	Primary commodity mined
SEAM_HEIGHT	Coal seam height in inches (coal mines only)
TYPE_OF_MINE	Type of mine
MINE_STATUS	Status of operation of mine
AVG_EMP_TOTAL	Average number of employees
EMP_HRS_TOTAL	Total number of employee hours
PCT_HRS_UNDERGROUND	Proportion of employee hours in underground operations
PCT_HRS_SURFACE	Proportion of employee hours at surface operations of underground mine
PCT_HRS_STRIP	Proportion of employee hours at strip mine
PCT_HRS_AUGER	Proportion of employee hours in auger mining
PCT_HRS_CULM_BANK	Proportion of employee hours in culm bank operations
PCT_HRS_DREDGE	Proportion of employee hours in dredge operations
PCT_HRS_OTHER_SURFACE	Proportion of employee hours in other surface mining operations
PCT_HRS_SHOP_YARD	Proportion of employee hours in independent shops and yards
PCT_HRS_MILL_PREP	Proportion of employee hours in mills or prep plants
PCT_HRS_OFFICE	Proportion of employee hours in offices
NUM_INJURIES	Total number of accidents reported

proximately the same number of observations from different years and that most mines are sand and gravel as well as surface mines.

Table 6 shows some summary statistics of the numerical variables. From the table, we see that the average number of employee ranges from 1 to 3115. The compositional components are those variables that start with PCT; each component refers to the proportion of employee hours spent in various types or methods of mining work such as strip-ping and auger mining. In the data, all compositional components contain zeros. In addition, most compositional components are skewed as the 3rd quantiles are zero. Table 7 shows the percentages of zeros of the compositional variables. From the table, we see that most values of many compositional variables are just zeros.

¹ The data set is available at <https://www.soa.org/globalassets/assets/files/edu/2018/2018-12-exam-pa-data-file.zip>.

Table 5. Summary of categorical variables.

YEAR	Number of observations	TYPE_OF_MINE	Number of observations
2013	13,759	Mill	2,578
2014	13,604	Sand and gravel	25,414
2015	13,294	Surface	23,091
2016	13,089	Underground	2,663

Table 6. Summary of numerical variables.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
AVG_EMP_TOTAL	1	3	5	17.8419	12	3,115
PCT_HRS_UNDERGROUND	0	0	0	0.0345	0	1
PCT_HRS_SURFACE	0	0	0	0.0087	0	1
PCT_HRS_STRIP	0	0.3299	0.8887	0.6801	1	1
PCT_HRS_AUGER	0	0	0	0.0046	0	1
PCT_HRS_CULM_BANK	0	0	0	0.0046	0	1
PCT_HRS_DREDGE	0	0	0	0.045	0	1
PCT_HRS_OTHER_SURFACE	0	0	0	0.0007	0	1
PCT_HRS_SHOP_YARD	0	0	0	0.0036	0	1
PCT_HRS_MILL_PREP	0	0	0	0.106	0	1
PCT_HRS_OFFICE	0	0	0.0312	0.1122	0.1685	1
NUM_INJURIES	0	0	0	0.4705	0	86

Table 7. Percentage of zeros of the compositional variables.

Variable	Percentage of zeros
PCT_HRS_UNDERGROUND	95.34
PCT_HRS_SURFACE	95.99
PCT_HRS_STRIP	18.42
PCT_HRS_AUGER	99.49
PCT_HRS_CULM_BANK	99.46
PCT_HRS_DREDGE	94.61
PCT_HRS_OTHER_SURFACE	99.91
PCT_HRS_SHOP_YARD	99.59
PCT_HRS_MILL_PREP	81.24
PCT_HRS_OFFICE	36.93

Figure 1(a) shows the proportions of employee hours in different categories from the first ten observations. Figure 1(b) shows the boxplots of the compositional variables. From the figures, we also see that the compositional variables have many zero values. Zeros in compositional data are irregular values. Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015) discussed several strategies to handle such irregular values. One strategy is to replace zeros by small positive values. In the mine dataset, we can treat zeros as rounded zeros below a detection limit. In this

paper, we replace zero components with 10^{-6} , which corresponds to a detection limit of one hour among one million employee hours.

The response variable is the total number of injuries reported. Figure 2 shows a frequency distribution of the response variable. The summary statistics given in Table 6 and Figure 2 show that the response variable is also highly skewed with many zeros. Figure 3 shows a scatter plot between the variable AVG_EMP_TOTAL and the response variable NUM_INJURIES. From the figure, we see that the two variables have a positive relationship. In our models, we will use the variable AVG_EMP_TOTAL as the exposure.

3.2. RESULTS

In this subsection, we present the results of fitting the four negative binomial regression models (see Table 3) to the mine data set. We follow the following procedure to fit and test the models NBILR, NBPCA, and NBEPCA:

1. Replace zeros with small positive values (i.e., 10^{-6}).
2. Transform the compositional variables for the whole mine data set. For the model NBILR, we transform the compositional variables with the ILR method. For the model NBPCA, we first transform the compositional variables with the CLR method and then apply the standard PCA to the transformed data. Both the ILR method and the CLR method are available in the

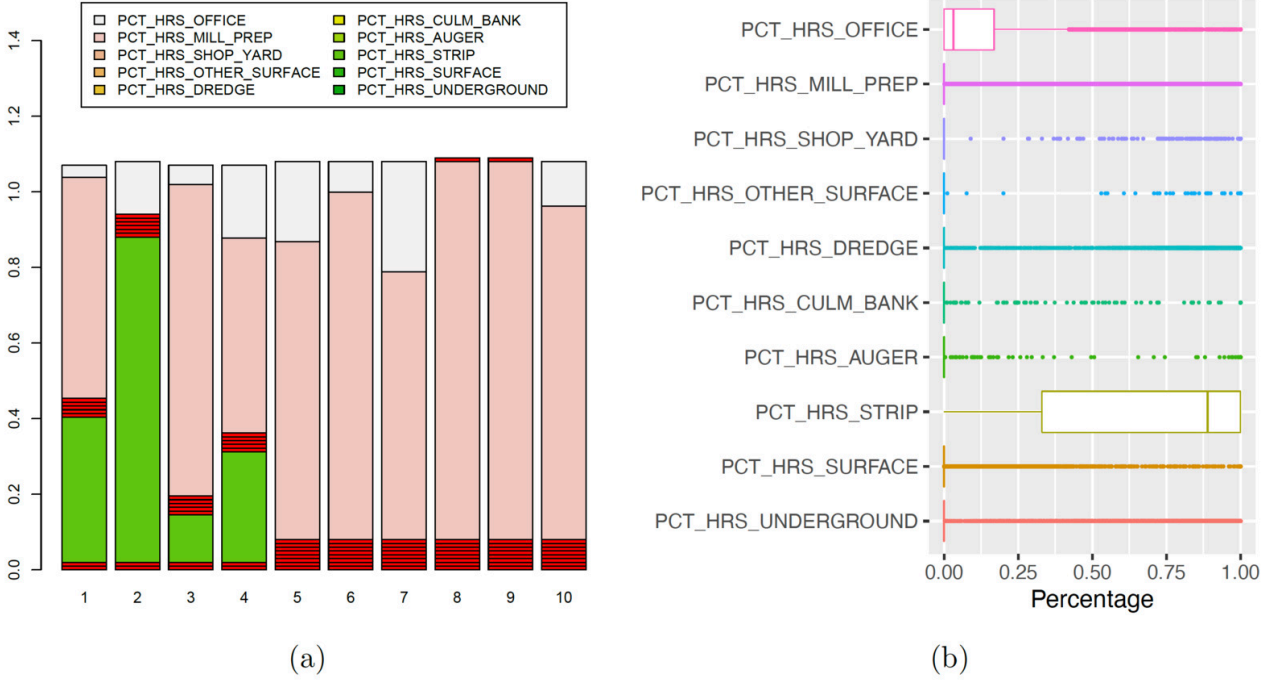


Figure 1. Compositional variables of the mine dataset.

The left figure shows the barplots of the compositional components of 10 observations. The small red bars represent zero proportions. The right figure shows the boxplots of the compositional covariates.

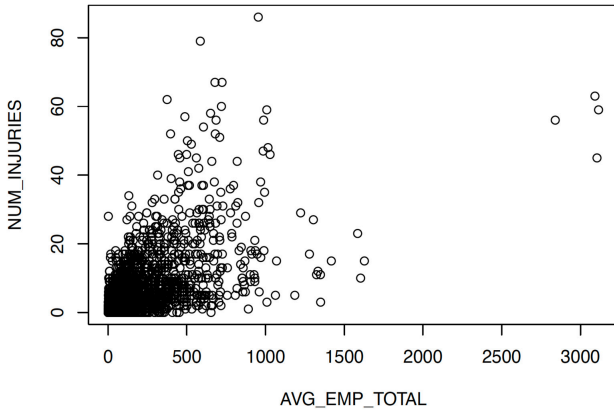


Figure 3. A scatter plot between the average number of employees and the total number of injuries reported.

R package *compositions*. For the model NBEPKA, we transform the compositional variable with the EPCA method implemented by Avalos et al. (2018).²

3. Split the transformed data into a training set and a test set. In particular, we use data from 2013 to 2015 as the training data and use data from 2016 as the test data.
4. Estimate parameters based on the training set.
5. Make predictions for the test set and calculate the out-of-sample validation measures.

We follow similar procedures to fit and test the model NB except that we do not transform the compositional variables.

We fitted the four models described in Table 3 to the mine dataset according to the afore-mentioned procedure. Table 8 shows the in-sample and out-of-sample chi-square statistics, as well as the mean prediction error (*ME*), produced by the four models. From the table, we see that the negative binomial model with EPCA transformed data performs the best among the four models. The model NBEPKA produced the lowest chi-square statistics for both the training data and the test data. A similar conclusion can be drawn based on the mean prediction error.

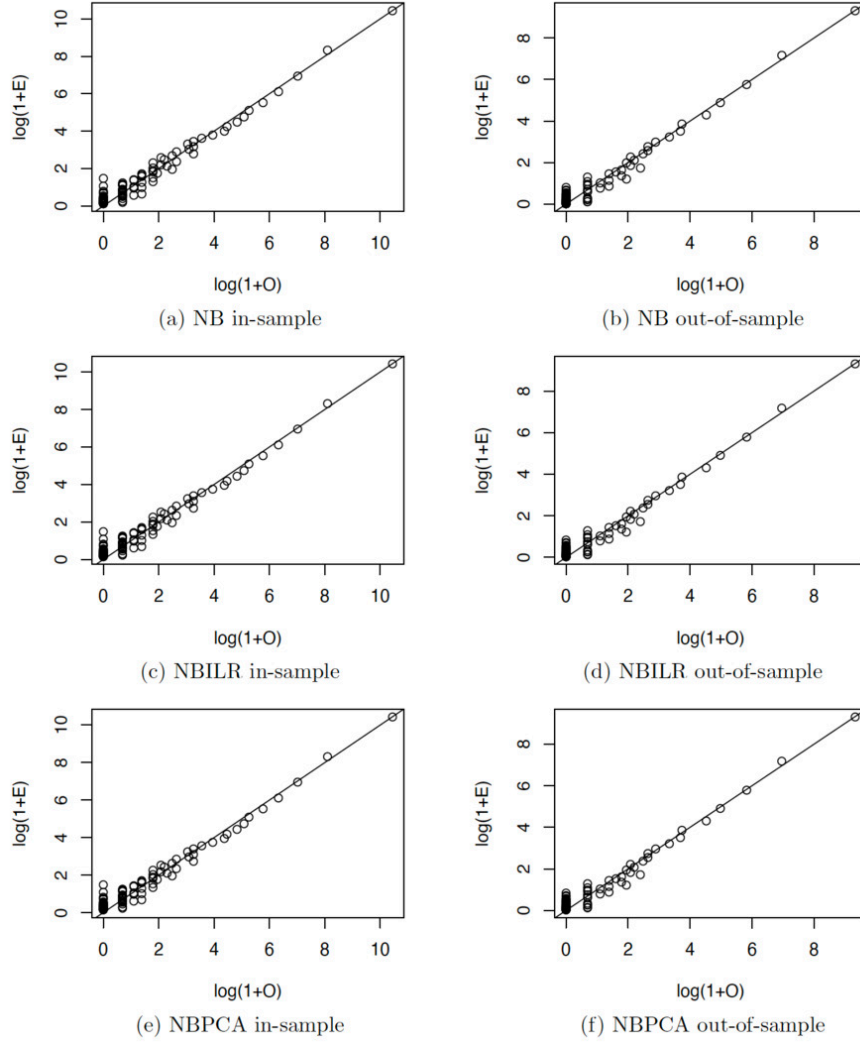
Figure 4 and Figure 5 show the scatter plots of the observed number of mines with different number of injuries and that predicted by the models at log scales. That is, the scatter plots show the relationship between $\log(1 + O_j)$ and $\log(1 + E_j)$ for $j = 0, 1, \dots, 99$. We use log scale here because the O_j 's and E_j 's have big differences for different j 's. From the scatter plots, we see that all the models predict the number of mines with different number of injuries quite well. We cannot tell the differences of the four models from the scatter plots. However, the chi-square statistics shown in Table 8 numerically demonstrates the superiority of the NBEPKA model.

For the NBPCA model, we selected the first five principal components produced by applying the standard PCA to the CLR-transformed data. Figure 6 shows the proportion of

² The R code of the EPCA method is available at <https://github.com/sistm/CoDa-PCA>.

Table 8. In-sample and out-of-sample validation measures produced by the four models.

Model	In-sample χ^2	Out-of-sample χ^2	In-sample mean prediction error	Out-of-sample mean prediction error
NB	316.0295	113.1281	-0.0155	-0.0295
NBILR	334.9093	113.1498	-0.0178	-0.0377
NBPCA	338.4336	114.2693	-0.0174	-0.0366
NBEPCA	303.9178	111.8998	-0.0024	-0.0246

**Figure 4. Scatter plots of the observed number of mines with different number of injuries and that predicted by models NB, NBILR, and NBPCA at log scales.**

variation explained by different numbers of principal components. The first five principal components can explain more than 95% of the variation of the original data. For the NBEPCA model, we selected the first five principal components for the same reason. The EPCA method used to transform the compositional variables is slow and took a normal desktop about 20 minutes to finish the computation. The reason is that the EPCA method we used is implemented in R, which is a scripting language that may not be efficient for iterative optimization.

Table 9 shows the estimated regression coefficients of the model NB. In this model, we used the original compositional variables by assuming that these variables contain absolute information. Since the sum of the compositional variables equals 1, we excluded the first compositional variable to avoid overparameterization. Table 9 shows the estimated regression coefficients for the nine compositional variables. The P values of these coefficients show that all the compositional variables are statistically significant.

Table 10 shows the estimated regression coefficients of the model NBILR. As shown: in Table 10(a), the ten compo-

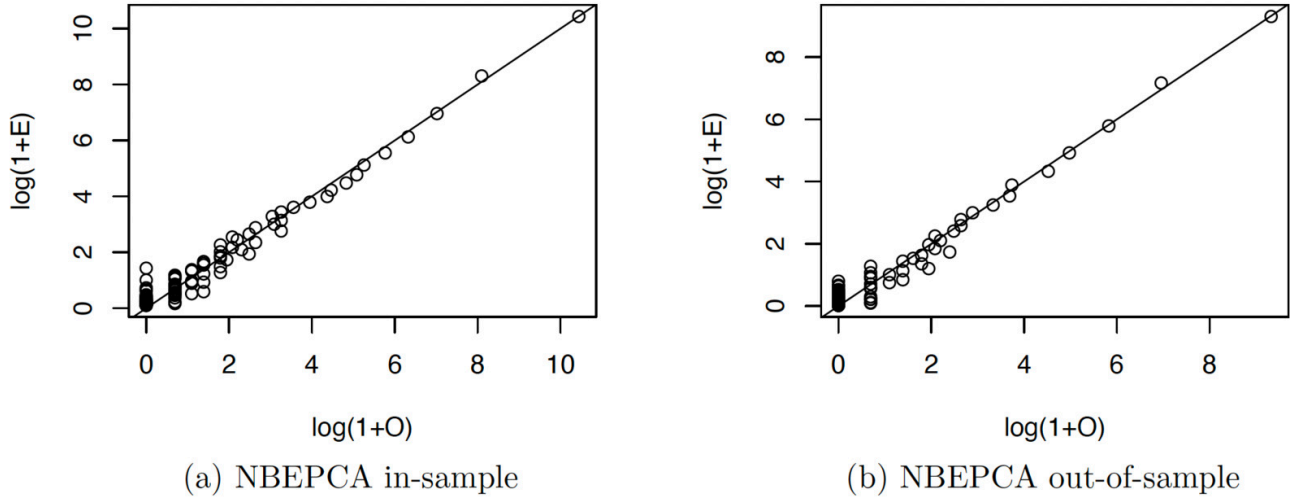


Figure 5. Scatter plots of the observed number of mines with different number of injuries and that predicted by the model NBEPCA at log scales.

Table 9. Estimated regression coefficients of the model NB.

Predictor	Coefficient	Standard Error	z value	P value
(Intercept)	-2.5173	0.1221	-20.6184	0
Sand and gravel	-0.0423	0.0649	-0.6521	0.5143
Surface	0.1602	0.0506	3.169	0.0015
Underground	-0.2895	0.1121	-2.582	0.0098
PCT_HRS_SURFACE	-1.4454	0.2305	-6.2709	0
PCT_HRS_STRIP	-1.5714	0.1238	-12.6966	0
PCT_HRS_AUGER	-1.182	0.2555	-4.6264	0
PCT_HRS_CULM_BANK	-1.6338	0.2886	-5.6615	0
PCT_HRS_DREDGE	-1.2183	0.147	-8.2891	0
PCT_HRS_OTHER_SURFACE	-1.2585	0.4022	-3.1286	0.0018
PCT_HRS_SHOP_YARD	-1.2796	0.2217	-5.7722	0
PCT_HRS_MILL_PREP	-0.9255	0.1205	-7.6801	0
PCT_HRS_OFFICE	-2.2877	0.1494	-15.3078	0

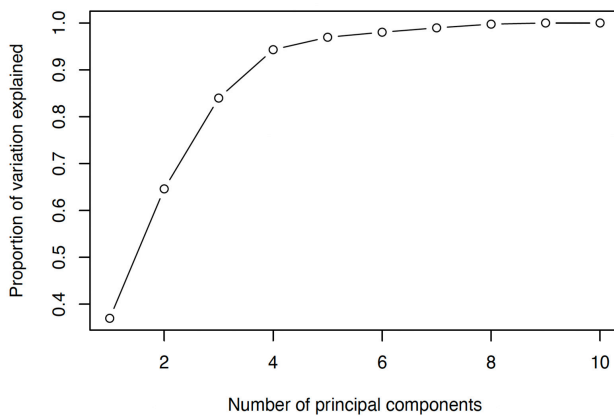


Figure 6. Proportion of variation explained with different principal components.

sitional variables are transformed to nine variables by the ILR method. Table 10(b) shows the regression coefficients transformed back to the CLR proportions by the inverse operation of the ILR method. From Table 10(b), we see that the coefficients of the ten compositional variables are similar and are quite different from the coefficients of the ILR variables V_1 to V_9 . From the P values shown in Table 10(a), we see that some of the ILR transformed variables are not significant. For example, the variable V_5 has a P value of 0.9652 and the variable V_7 has a P value of 0.1649. Both variables have a P value greater than 0.1. This suggests that it is appropriate to reduce the dimensionality of the compositional variables.

The regression coefficients of the ILR transformed variables V_1 to V_9 shown in Table 10 vary a lot. It is challenging to interpret those coefficients. From the P values shown in Table 10(a), we see that some of the ILR transformed variables are not significant. For example, the variable V_5

(a)

Predictor	Coefficient	Std. Error	z value	P value
(Intercept)	-3.7525	0.0505	-74.3196	0
Sand & gravel	-0.2356	0.0795	-2.9651	0.003
Surface	0.0444	0.0689	0.6446	0.5192
Underground	-0.7147	0.1614	-4.4288	0
V1	-0.0536	0.0111	-4.8275	0
V2	-0.053	0.0061	-8.6443	0
V3	-0.0248	0.0127	-1.9471	0.0515
V4	-0.0674	0.0168	-4.0039	0.0001
V5	0.0003	0.007	0.0437	0.9652
V6	-0.044	0.0224	-1.9631	0.0496
V7	-0.018	0.013	-1.3886	0.1649
V8	0.0143	0.0025	5.6681	0
V9	-0.0057	0.0028	-2.0133	0.0441

(b)

Predictor	Coefficient
PCT_HRS_UNDERGROUND	0.1093
PCT_HRS_SURFACE	0.1013
PCT_HRS_STRIP	0.0986
PCT_HRS_AUGER	0.1001
PCT_HRS_CULMBANK	0.0949
PCT_HRS_DREDGE	0.1008
PCT_HRS_OTHER_SURFACE	0.0961
PCT_HRS_SHOP_YARD	0.0982
PCT_HRS_MILL_PREP	0.1013
PCT_HRS_OFFICE	0.0994

Table 10. Estimated regression coefficients of the model NBILR. Table (a) shows the coefficients of ILR transformed data. Table (b) shows the coefficients transformed back to the CLR proportions.

has a P value of 0.9652 and the variable V7 has a P value of 0.1649. Both variables have a P value greater than 0.1. This suggests that it is appropriate to reduce the dimensionality of the compositional variables.

Table 11 shows the estimated regression coefficients of the models NBPCA and NBEPKA. We used five principal components in both models. From Table 11(b), we see that all principal components produced by the EPCA method are significant. However, Table 11(a) shows that the third principal component PC3 produced by the traditional PCA method is not significant as it has a P value of 0.8214.

In summary, our numerical results show that the EPCA method is able to produce better principal components than the traditional PCA method and that using the principal components produced by the EPCA method can improve the prediction accuracy of the negative binomial models with compositional covariates.

4. CONCLUSIONS

It is important to recognize the unique aspects of compositional data when using and fitting them as feature variables in predictive models. Compositional data are commonly presented as vectors of proportions, percentages, concentrations, or frequencies. A peculiarity of these vectors is that their sum is constrained to be some fixed constant, e.g., 100%. Due to such constraints, compositional

(a) NBPCA

Predictor	Coefficient	Std. Error	z value	P value
(Intercept)	-3.8635	0.0655	-58.9829	0
Sand & gravel	-0.2261	0.0789	-2.8656	0.0042
Surface	0.0588	0.0681	0.8631	0.3881
Underground	-0.3017	0.1319	-2.2876	0.0222
PC1	0.0274	0.0033	8.3922	0
PC2	0.0245	0.0035	7.0887	0
PC3	-0.0007	0.0031	-0.2257	0.8214
PC4	-0.0356	0.0071	-4.9989	0
PC5	-0.0514	0.0096	-5.3658	0

(b) NBEPKA

Predictor	Coefficient	Std. Error	z value	P value
(Intercept)	-3.2533	0.1209	-26.9052	0
Sand & gravel	-0.348	0.0719	-4.8394	0
Surface	-0.1203	0.0618	-1.9444	0.0518
Underground	-0.1894	0.1153	-1.6428	0.1004
PC1	0.156	0.0184	8.479	0
PC2	0.0943	0.0073	12.8948	0
PC3	-0.0414	0.0124	-3.329	0.0009
PC4	0.0314	0.0152	2.0676	0.0387
PC5	-0.2678	0.027	-9.9187	0

Table 11. Estimated regression coefficients of the models NBPCA and NBEPKA.

data have the following properties: they carry relative information (scale invariance); equivalent results should be yielded when the ordering of the parts in the composition is changed (permutation invariance); and results should remain consistent if a noninformative part is removed (subcomposition coherence).

In this article, we offer methods to handle regression models with compositional covariates using a mine data set for empirical investigation. We built negative binomial regression models to predict the number of injuries from different mines. In particular, we built four negative binomial regression models: a model with the original compositional components, a model with ILR-transformed compositional variables, a model with principal components produced by the traditional PCA, and a model with principal components produced by the EPCA. Our numerical results show that the EPCA method for controlling compositional covariates produces principal components that are significant predictors and improve the prediction accuracy of the regression model. However, the EPCA method may unavoidably have some limitations. For example, interpreting the regression coefficients is not straightforward due to the transformation. In addition, the log transformation requires dealing with zeros in the data.

Submitted: September 19, 2022 EDT. Accepted: January 20, 2023 EDT. Published: May 14, 2024 EDT.

REFERENCES

- Aitchison, J. 1994. "Principles of Compositional Data Analysis." *Multivariate Analysis and Its Applications* 24:73–81. <https://doi.org/10.1214/lnms/1215463786>.
- Aitchison, J., and J. J. Egozcue. 2005. "Compositional Data Analysis: Where Are We and Where Should We Be Heading?" *Mathematical Geology* 37 (7): 829–50. <https://doi.org/10.1007/s11004-005-7383-7>.
- Aitchison, J. 2003. *The Statistical Analysis of Compositional Data*. Caldwell, NJ: Blackburn.
- Avalos, M., R. Nock, C.S. Ong, J. Rouar, and K. Sun. 2018. "Representation Learning of Compositional Data." In *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 31:6679–89. Red Hook, NY: Curran Associates.
- Belles-Sampera, J., M. Guillen, and M. Santolino. 2016. "Compositional Methods Applied to Capital Allocation Problems." *Journal of Risk* 19 (2): 15–30. <https://doi.org/10.21314/jor.2016.345>.
- Boonen, T. J., M. Guillen, and M. Santolino. 2019. "Forecasting Compositional Risk Allocations." *Insurance: Mathematics and Economics* 84 (January):79–86. <https://doi.org/10.1016/j.insmatheco.2018.10.002>.
- Collins, M., S. Dasgupta, and R. E. Schapire. 2002. "A Generalization of Principal Component Analysis to the Exponential Family." In *Advances in Neural Information Processing Systems 14*, edited by T. Dietterich, S. Becker, and Z. Ghahramani, 14:617–24. Cambridge, MA: The MIT Press. <https://doi.org/10.7551/mitpress/1120.003.0084>.
- Denuit, M., M. Guillen, and J. Trufin. 2019. "Multivariate Credibility Modelling for Usage-Based Motor Insurance Pricing with Behavioural Data." *Annals of Actuarial Science* 13 (2): 378–99. <https://doi.org/10.1017/s1748499518000349>.
- Frees, E. W. 2009. *Regression Modeling with Actuarial and Financial Applications*. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/cbo9780511814372>.
- Gan, G., and E. Valdez. 2018. *Actuarial Statistics with R: Theory and Case Studies*. New Hartford, CT: ACTEX Learning.
- Guillen, M., J. P. Nielsen, M. Ayuso, and A. M. Pérez-Marín. 2019. "The Use of Telematics Devices to Improve Automobile Insurance Rates." *Risk Analysis* 39 (3): 662–72. <https://doi.org/10.1111/risa.13172>.
- Guillen, M., J. P. Nielsen, A. M. Pérez-Marín, and V. Elpidorou. 2020. "Can Automobile Insurance Telematics Predict the Risk of Near-Miss Events?" *North American Actuarial Journal* 24 (1): 141–52. <https://doi.org/10.1080/10920277.2019.1627221>.
- Kjærgaard, S., Y. E. Ergemen, M. Kallestrup-Lamb, J. Oeppen, and R. Lindahl-Jacobsen. 2019. "Forecasting Causes of Death by Using Compositional Data Analysis: The Case of Cancer Deaths." *Journal of the Royal Statistical Society Series C* 68 (5): 1351–70. <https://doi.org/10.1111/rssc.12357>.
- Pawlowsky-Glahn, V., and A. Buccianti. 2011. *Compositional Data Analysis: Theory and Applications*. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9781119976462>.
- Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado. 2015. *Modelling and Analysis of Compositional Data*. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9781119003144>.
- Pearson, K. 1897. "Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs." *Proceedings of the Royal Society of London* 60 (359–367): 489–98. <https://doi.org/10.1098/rspl.1896.0076>.
- Pesantez-Narvaez, J., M. Guillen, and M. Alcañiz. 2019. "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression." *Risks* 7 (2): 70. <https://doi.org/10.3390/risks7020070>.
- So, B., J.-P. Boucher, and E. A. Valdez. 2021. "Cost-Sensitive Multi-Class AdaBoost for Understanding Driving Behavior Based on Telematics." *ASTIN Bulletin: The Journal of the International Actuarial Association* 51 (3): 719–51. <https://doi.org/10.1017/asb.2021.22>.
- Tipping, M. E., and C. M. Bishop. 1999. "Probabilistic Principal Component Analysis." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61 (3): 611–22. <https://doi.org/10.1111/1467-9868.00196>.

Verbelen, R., K. Antonio, and G. Claeskens. 2018.
“Unravelling the Predictive Power of Telematics Data
in Car Insurance Pricing.” *Journal of the Royal
Statistical Society Series C: Applied Statistics* 67 (5):
1275–1304. <https://doi.org/10.1111/rssc.12283>.